

## PATENT ABSTRACTS OF JAPAN

(11)Publication number : 08-329079

(43)Date of publication of application : 13.12.1996

(51)Int.Cl. G06F 17/24  
G06F 17/27  
G06F 17/30

(21)Application number : 07-161398

(71)Applicant : HITACHI LTD

(22)Date of filing : 05.06.1995

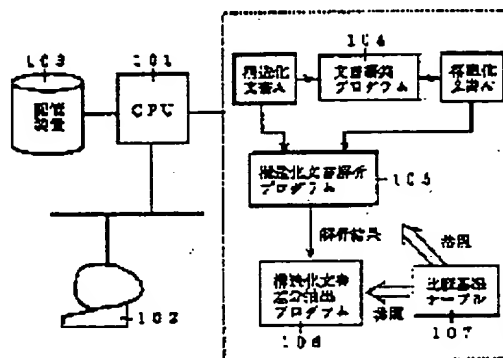
(72)Inventor : AOYAMA YUKI  
TONO JUNICHI

## (54) METHOD AND DEVICE FOR EXTRACTING STRUCTURED DOCUMENT DIFFERENCE

## (57)Abstract:

PURPOSE: To extract proper differences of a structure document as a document editor feels in consideration of the logical meaning and structure of the structured document.

CONSTITUTION: A document editing program 104 edits the structured document and stores it in a storage device, a structured document program 105 analyzes the logical structures of respective structured documents before and after editing which are read out of the stored device by referring to a comparison reference 107 set for the logical structures of the structured documents before and after editing, and a structured document difference extracting program 106 extracts differences between the structured documents according to the analysis result so that the comparison reference 107 is met. The comparison reference 107 is a table consisting of tags indicating the logical structures and the kinds of the reference to the tags, and as kinds of the reference, there are (1) a tag for comparing contents only when the tag itself is coincident, (2) a tag for ignoring differences of contents of the tag at the time of comparison, (3) a group of tags which are identical in logical meaning, and (4) a group of tags which do not compare contents.



## LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision  
of rejection]

[Date of requesting appeal against examiner's  
decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2000 Japanese Patent Office

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平8-329079

(43) 公開日 平成8年(1996)12月13日

(51) Int.Cl. <sup>8</sup>	識別記号	序内整理番号	F I	技術表示箇所
G 0 6 F 17/24		9288-5L	G 0 6 F 15/20	5 5 4 N
17/27		9288-5L		5 5 0 F
17/30		9194-5L	15/403	3 5 0 A

審査請求 未請求 請求項の数10 F D (全 12 頁)

(21) 出願番号 特願平7-161398

(22) 出願日 平成7年(1995)6月5日

(71) 出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72) 発明者 青山 ゆき

神奈川県川崎市麻生区王禅寺1099番地 株

式会社日立製作所システム開発研究所内

(72) 発明者 東野 純一

神奈川県川崎市麻生区王禅寺1099番地 株

式会社日立製作所システム開発研究所内

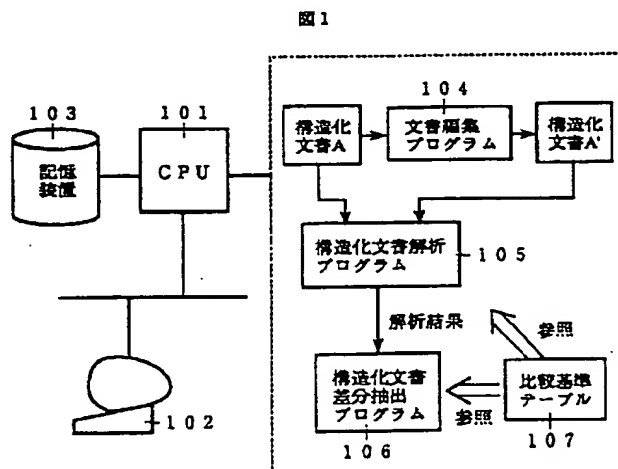
(74) 代理人 弁理士 笹岡 茂 (外1名)

(54) 【発明の名称】 構造化文書差分抽出方法および装置

(57) 【要約】

【目的】 構造化文書の論理的な意味や構造を考慮し、文書編集者の感覚に合った適切な構造化文書の差分を抽出することにある。

【構成】 文書編集プログラム104により構造化文書を文書編集して記憶装置に記憶し、編集前後の各構造化文書の論理構造に対して設定された比較基準107を参照して、記憶装置から読み出した編集前後の各構造化文書の論理構造を構造化文書解析プログラム105で解析し、この結果に従い比較基準107を満たすように構造化文書間の差分を構造化文書差分抽出プログラム106により抽出する。比較基準107を、論理構造を示すタグと該タグに対する基準の種類からなるテーブルとし、基準の種類を、(1)タグ自身が一致したときのみその中身を比較するタグ、(2)比較する際、そのタグの中身の差異を無視するタグ、(3)論理的な意味として同じタグの組、(4)中身を比較しないタグの組、としている。



## 【特許請求の範囲】

【請求項 1】 記憶装置と処理装置を備え、前記記憶装置に削除、挿入、または変更などの編集を実施する前後の構造化文書を記憶し、前記処理装置により前記編集前後の両構造化文書の一致しない文字列を差分として抽出する構造化文書差分抽出方法において、  
構造化文書を文書編集して前記記憶装置に記憶し、前記記憶装置から読み出した編集前後の各構造化文書の論理構造を、編集前後の各構造化文書の論理構造に対して設定された比較基準に基づき構造化文書解析し、該構造化文書解析の結果に従い、前記比較基準を満たすよう構造化文書間の差分を抽出することを特徴とする構造化文書差分抽出方法。

【請求項 2】 請求項 1 記載の構造化文書差分抽出方法において、  
前記比較基準を、論理構造を示すタグと該タグに対する基準の種類からなるテーブルとすることを特徴とする構造化文書差分抽出方法。

【請求項 3】 請求項 2 記載の構造化文書差分抽出方法において、  
前記比較基準におけるタグに対する基準の種類として、  
(1) タグ自身が一致したときのみその中身を比較するタグ、  
(2) 比較する際、そのタグの中身の差異を無視するタグ、  
(3) 論理的な意味として同じタグの組、  
(4) 中身を比較しないタグの組、の 4 つの基準の種類を定義しておくことを特徴とする構造化文書差分抽出方法。

【請求項 4】 請求項 1 乃至請求項 3 のいずれかの請求項記載の構造化文書差分抽出方法において、  
前記構造化文書解析により文書の構造を表わす文書木を作成し、該文書木のノード単位で構造化文書間の差分の抽出を行ない、一致しなかったノード同士に対して文字単位で差分を抽出することを特徴とする構造化文書差分抽出方法。

【請求項 5】 請求項 4 記載の構造化文書差分抽出方法において、  
前記構造化文書解析によって文書の構造を表す文書木を作成する際に、前記比較基準に応じて、文書木のノードの割り当て方法を変えることを特徴とする構造化文書差分抽出方法。

【請求項 6】 記憶装置と処理装置を備え、前記記憶装置に削除、挿入、または変更などの編集を実施する前後の構造化文書を記憶し、前記処理装置により前記編集前後の両構造化文書の一致しない文字列を差分として抽出する構造化文書差分抽出装置において、  
前記処理装置は、  
構造化文書を編集して前記記憶装置に記憶する文書編集手段と、

前記記憶装置から読み出した編集前後の各構造化文書の論理構造を、編集前後の各構造化文書の論理構造に対して設定された比較基準に基づき構造化文書解析する構造化文書解析手段と、

該構造化文書解析の結果に従い、前記比較基準を満たすよう構造化文書間の差分を抽出する構造化文書差分抽出手段を備えることを特徴とする構造化文書差分抽出装置。

【請求項 7】 請求項 6 記載の構造化文書差分抽出装置において、  
前記比較基準を、論理構造を示すタグと該タグに対する基準の種類からなるテーブルとすることを特徴とする構造化文書差分抽出装置。

【請求項 8】 請求項 7 記載の構造化文書差分抽出装置において、  
前記比較基準におけるタグに対する基準の種類として、  
(1) タグ自身が一致したときのみその中身を比較するタグ、  
(2) 比較する際、そのタグの中身の差異を無視するタグ、  
(3) 論理的な意味として同じタグの組、  
(4) 中身を比較しないタグの組、の 4 つの基準の種類を定義しておくことを特徴とする構造化文書差分抽出装置。

【請求項 9】 請求項 6 乃至請求項 8 のいずれかの請求項記載の構造化文書差分抽出装置において、  
前記構造化文書解析手段は文書の構造を表わす文書木を作成し、前記構造化文書差分抽出手段は作成された文書木のノード単位で構造化文書間の差分の抽出を行ない、一致しなかったノード同士に対して文字単位で差分を抽出することを特徴とする構造化文書差分抽出装置。

【請求項 10】 請求項 9 記載の構造化文書差分抽出装置において、  
前記構造化文書解析手段は文書の構造を表す文書木を作成する際に、前記比較基準に応じて、文書木のノードの割り当てを変更することを特徴とする構造化文書差分抽出装置。

## 【発明の詳細な説明】

【0001】

【産業上の利用分野】 本発明は、電子ファイルとして記憶されている構造化文書間の差分文字列を抽出することができるワープロ等の文書処理装置における構造化文書差分文字列抽出方法及び装置に関する。

【0002】

【従来の技術】 構造化文書とは、文書の論理的な構造の情報、例えば“文書中のこの部分は章である”、“この部分はタイトルである”といった情報が埋め込まれた文書のことである。また、文書間の差分抽出とは、文書を構成する段落、行、文字等の要素を単位に、これらの要素がもっともよく一致する組合せを検出し、一致しない

要素を差分として抽出することである。例えば、差分抽出の対象とする二つの文書を「A B C D E F G」と「A C D A E F H」とした場合、要素をA, B, C, D, E, F, G, Hとして二つの文書を要素単位で比較した時、もっともよく一致する組合せとして”A C D E Fが対応する”と検出し、また差分として”Bが削除”、”AがDの後に挿入”、”GがHに変更”と抽出することである。従来の差分抽出方式には、特開平2-255964号公報などがあり、句読点、行、単語、文字などを単位に比較を行っている。この方式を構造化文書に適用すると、文書中に埋め込まれた論理的な構造を表す文字列も、文書中の他の文字列と同様に比較を行う。

【0003】

【発明が解決しようとする課題】しかしながら、構造化文書を通常の文書と同様の手段で差分抽出した場合、結果が文書の論理構造と合わず文書編集者にとって適切でない場合がある。次に例を示し説明する。

【0004】（例1）差分抽出において文書の論理構造が合わないもの同士を対応付けてしまい、文書編集者にとって適切でない抽出結果となる場合を、図3の構造化文書を例にとり説明する。図3の構造化文書はSGML (Standard Generalized Markup Language) で記述されたもので、〈A〉と〈／A〉で挟まれた文字列が、論理構造Aに属していることを意味する。すなわち、図3(a)の〈氏名〉と〈／氏名〉で挟まれた文字列“平成太郎”が、論理構造“氏名”に属する。また、この論理構造を表すマークのことをタグと呼び、〈A〉と〈／A〉はそれぞれ開始タグ、終了タグと呼ぶ。従来の手法により、図3の

(a), (b)二つの構造化文書の差分文字列を抽出した結果を図4に示す。図4(b)は、図3(a)の構造化文書を基準として図3(b)の構造化文書との差分を取った場合の差分文字列の抽出結果であり、図4(a)は、図3(b)の構造化文書を基準として図3(a)の構造化文書との差分を取った場合の差分文字列の抽出結果である。図4を見ると、〈氏名〉の“平成”と〈発信日〉の“平成”が差分として抽出されていない。これは、“平成”同士が一致し、対応付けられてしまったことによる。しかし、この論理構造の合わない“平成”の対応付けは、文書編集者にとって意味がないことは明らかである。

【0005】（例2）文書の構造の挿入が起きたために、差分抽出において文書の構造にまたがって文字列を対応付けてしまい、文書編集者にとって適切でない抽出結果となる場合を、図5の構造化文書を例にとり説明する。図5は、(a)の第1章の前に、章を一つ挿入したものが(b)となっている。従来の手法により、図5の(a), (b)二つの構造化文書の差分文字列を抽出した例を図6示す。図6は図4の場合と同様であり、図6(b)が図5(a)を基準として図5(b)との差分を

取った場合の差分文字列の抽出結果であり、図6(a)が図5(b)を基準として図5(a)との差分を取った場合の差分文字列の抽出結果である。図6を見ると、

(a)の第1章は(b)の第2章と同じであるにもかかわらず、(a)の第1章が、(b)の第1章と第2章にまたがって対応している。これも、文書編集者に対しては適切でない。ここで、図5(a)における「構造化文書」と同じ文字列が図5(b)には2度現われているため、図6(b)では、最初の「構造化文書」は一致とされ、2度目の「構造化文書」は不一致とされ、差分として抽出される。このことは、以下の差分抽出において共通した取扱いである。

【0006】（例3）文書の論理的な意味は同じであるのに、論理構造を表すマークが異なるためその中身同士が対応付けられず、文書編集者にとって適切でない抽出結果となる場合を、図7の構造化文書を例にとり説明する。図7では、文書の論理的な意味は項目であるのに、最初に出てくる項目だけ〈初項目〉というタグを付けている。従来の手法により、図7の(a), (b)二つの構造化文書の差分文字列を抽出した例を図8示す。図8は図4の場合と同様であり、図8(b)が図7(a)を基準として図7(b)との差分を取った場合の差分文字列の抽出結果であり、図8(a)が図7(b)を基準として図7(a)との差分を取った場合の差分文字列の抽出結果である。図8を見ると、〈初項目〉同士が対応付けられ、その中身の文字列が比較されていることが分かる。文書編集者にとっては〈初項目〉と〈項目〉の論理的な意味は等しく、タグの中身を優先して対応させるべきである。そこで、構造化文書間の差分を抽出する場合、構造化文書の論理的な意味や構造を考慮した比較が必要となるが、従来の方式では、論理的な構造を表す文字列も文書中の他の文字列と同様に比較を行うため、実現できなかった。

【0007】本発明の目的は、構造化文書の論理的な意味や構造を考慮し、文書編集者の感覚に合った適切な構造化文書の差分を抽出することにある。

【0008】

【課題を解決するための手段】上記目的を達成するため、本発明は、記憶装置と処理装置を備え、前記記憶装置に削除、挿入、または変更などの編集を実施する前後の構造化文書を記憶し、前記処理装置により前記編集前後の両構造化文書の一致しない文字列を差分として抽出する構造化文書差分抽出方法において、構造化文書を文書編集して前記記憶装置に記憶し、前記記憶装置から読み出した編集前後の各構造化文書の論理構造を、編集前後の各構造化文書の論理構造に対して設定された比較基準に基づき構造化文書解析し、該構造化文書解析の結果に従い、前記比較基準を満たすよう構造化文書間の差分を抽出するようにしている。前記比較基準を、論理構造を示すタグと該タグに対する基準の種類からなるテーブル

ルとするようにしている。さらに、前記比較基準におけるタグに対する基準の種類として、(1) タグ自身が一致したときのみその中身を比較するタグ、(2) 比較する際、そのタグの中身の差異を無視するタグ、(3) 論理的な意味として同じタグの組、(4) 中身を比較しないタグの組、の4つの基準の種類を定義しておくようにしている。さらに、前記構造化文書解析により文書の構造を表わす文書木を作成し、該文書木のノード単位で構造化文書間の差分の抽出を行ない、一致しなかったノード同士に対して文字単位で差分を抽出するようにしている。さらに、前記構造化文書解析によって文書の構造を表す文書木を作成する際に、前記比較基準に応じて、文書木のノードの割り当て方法を変えるようにしている。また、記憶装置と処理装置を備え、前記記憶装置に削除、挿入、または変更などの編集を実施する前後の構造化文書を記憶し、前記処理装置により前記編集前後の両構造化文書の一致しない文字列を差分として抽出する構造化文書差分抽出装置において、前記処理装置は、構造化文書を編集して前記記憶装置に記憶する文書編集手段と、前記記憶装置から読み出した編集前後の各構造化文書の論理構造を、編集前後の各構造化文書の論理構造に対して設定された比較基準に基づき構造化文書解析する構造化文書解析手段と、該構造化文書解析の結果に従い、前記比較基準を満たすよう構造化文書間の差分を抽出する構造化文書差分抽出手段を備えるようにしている。前記比較基準を、論理構造を示すタグと該タグに対する基準の種類からなるテーブルとするようにしている。さらに、前記比較基準におけるタグに対する基準の種類として、(1) タグ自身が一致したときのみその中身を比較するタグ、(2) 比較する際、そのタグの中身の差異を無視するタグ、(3) 論理的な意味として同じタグの組、(4) 中身を比較しないタグの組、の4つの基準の種類を定義しておくようにしている。さらに、前記構造化文書解析手段は文書の構造を表わす文書木を作成し、前記構造化文書差分抽出手段は作成された文書木のノード単位で構造化文書間の差分の抽出を行ない、一致しなかったノード同士に対して文字単位で差分を抽出するようにしている。さらに、前記構造化文書解析手段は文書の構造を表す文書木を作成する際に、前記比較基準に応じて、文書木のノードの割り当てを変更するようにしている。

#### 【0009】

【作用】上記手段により、本発明においては、構造化文書を編集し、編集された構造化文書の論理構造を構造化文書解析装置で解析し、その構造に応じて差分抽出の際の比較基準を設け、比較基準を満たすように差分文字列を抽出するので、論理構造に応じた、編集者の感覚に合う差分が抽出される。また、文書木のノード単位で差分抽出を行い、一致しなかったノード同士を文字単位で差分を抽出することで、構造にまたがった差分も抽出され

ない。

#### 【0010】

【実施例】以下、本発明の実施例を説明する。本実施例の構成を図1に示す。図1において、101はCPU、102は端末装置、103は文書を記憶するための記憶装置であり、CPU101には、文書の編集を行う文書編集プログラム104と、構造化文書を木構造に変換する構造化文書解析プログラム105と、構造化文書間の一致しない部分を差分として抽出する構造化文書差分抽出プログラム106と、差分抽出での比較基準を格納する比較基準テーブル107が設定されている。本実施例は、構造化文書としてSGML文書を例にとる。SGMLは、マーク付けされた構造化文書としてISOの世界標準として定められた文書記述言語のことである。また、SGML文書はDTD（文書型定義）によって、その論理構造が予め定義される。

【0011】本実施例の具体的な処理手順を、図2のフローチャートを用いて説明する。

手順201：文書編集プログラム104で、構造化文書の編集を行う。

手順202：比較対象であるSGML文書のDTDに対応した、比較基準テーブル107を読み込む。対応する比較基準テーブルが存在しない場合、テーブルの作成及び登録を行う。この比較基準テーブルは、次の4つの比較基準に該当するタグのテーブルである。

(1) 恒等タグ：タグ自身が一致したときのみ、その中身（開始タグと終了タグの間に挟まれる文字）を比較するタグである。

(2) 無視タグ：比較する際、そのタグの中身の差異を無視するタグである。

(3) 同等タグ：論理的な意味として同じタグの組である。

(4) 比較禁止タグ：中身を比較しないタグの組である。

【0012】手順203：差分抽出プログラム106が呼び出されたら、比較基準テーブル107を参照しながら、構造化文書を構造化文書解析プログラム105によって解析し、文書木を作成する。このとき、文書木の各ノードに割り当てる要素は次のルールを用いて行う。

(ルール1)：タグは1つのノードに割り当てる。

(ルール2)：開始タグと終了タグの間に挟まれた文字列は、開始タグの子ノードに割り当てる。

(ルール3)：終了タグは、開始タグの子ノードに割り当てる。

(ルール4)：恒等タグで挟まれた文字列は、開始タグ、終了タグを含めて1つのノードに割り当てる。

(ルール5)：無視タグおよび無視タグで挟まれた文字列は、ノードに割り当てない。

(ルール6)：同等タグは、同じタグ名に変換して、ノードに割り当てる。

【0013】手順204：文書木のノードを単位に差分抽出を行う。このとき、比較するタグ同士が比較禁止タグであれば、そのノード以下（子ノード）は比較しない。

手順205：一致しなかったノードのみ、今度は文字単位で差分抽出を行う。ただし、恒等タグのノードはノードの先頭文字であるタグが一致した場合のみ、文字単位の比較を行う。手順204で比較しなかった無視タグもこの段階で比較を行う。

手順206：端末装置102に差分結果の表示を行う。

【0014】（処理例1）実施例の具体的な処理例として、恒等タグをもつ場合を図3の文書例で説明する。

手順201：文書編集プログラム104で、構造化文書の編集を行う。図3の（a）から図3の（b）を編集したとする。

手順202：比較対象であるSGML文書のDTDに対応した比較基準テーブル107を読み込む。対応する比較基準テーブルが存在しない場合、テーブルの作成及び登録を行う。図3からは、例えば図9のような比較基準テーブルを作成する。すなわち、＜氏名＞および＜発信日＞を恒等タグとして定義し、タグ同士が一致しない限り、文字列同士を対応させないという意味を持つ。

【0015】手順203：差分抽出プログラム106が呼び出されたら、比較基準テーブル107を参照しながら、構造化文書を構造化文書解析プログラム105によって解析し、文書木を作成する。実施例で説明したルールを適用すると、図3の文書（a）、（b）から、図9の比較基準テーブルを参照することにより、図10の文書木（a）、（b）ができる。図10中の1001、1002は、（ルール4）によって、タグと中身の文字列が合わせて1つのノードに割り当てられている。

【0016】手順204：文書木のノードを単位に差分抽出を行う。ノードを単位に比較を行うため、恒等タグである＜氏名＞および＜発信日＞は、タグと中身の文字列が両者とも一致しない限り、対応付けられることはない。この場合、タグが一致しないため、タグおよびその中身が差分として抽出される。

手順205：一致しなかったノードのみ、今度は文字単位で差分抽出を行う。ただし、恒等タグのノードはノードの先頭文字であるタグが一致した場合のみ、文字単位の比較を行う。

【0017】手順206：端末装置102に差分結果の表示を行う。図3の文書（a）との文書（b）の差分抽出を行った結果例を図11に示す。図11（b）は、図3（a）の構造化文書を基準として図3（b）の構造化文書との差分を取った場合の差分文字列の抽出結果であり、図11（a）は、図3（b）の構造化文書を基準として図3（a）の構造化文書との差分を取った場合の差分文字列の抽出結果である。図11（b）では、ノード1001とノード1002におけるタグ（記号）とタグ

（発信日）が一致しないので、ノード1002全体の「（発信日）平成6年11月二十日（／発信日）」が差分として抽出され、また、図3（a）には図3（b）における「お元気ですか」の記載が無いので、「お元気ですか」が差分として抽出される。

【0018】以上の手順により差分抽出を行うと、タグが一致しないと中身を比較しても意味のないものを恒等タグとして登録しておけば、文書の論理構造が合わないもの同士を対応付けることがなくなり、編集者に対して、より適切な差分抽出結果を提示することが出来る。

【0019】（処理例2）実施例の具体的な処理例の2番目として、恒等タグおよび無視タグをもつ場合、および構造のずれが起きている場合を図5の文書例で説明する。

手順201：文書編集プログラム104で、構造化文書の編集を行う。図5の（a）から図5の（b）を編集したとする。

【0020】手順202：比較対象であるSGML文書のDTDに対応した、比較基準テーブル107を読み込む。対応する比較基準テーブルが存在しない場合、テーブルの作成及び登録を行う。図5の例では、例えば、図12のような比較基準テーブルを作成する。すなわち、＜著者名＞を恒等タグとして定義する。この場合、前述したように、タグ同士が一致した場合のみ、文字列同士を比較する。また、＜章番号＞を無視タグとして定義する。この場合、章番号の違いは無視する（差分抽出に影響を与えない）。

【0021】手順203：差分抽出プログラム106が呼び出されたら、比較基準テーブル107を参照しながら、SGML文書を構造化文書解析プログラム105によって解析し、文書木を作成する。実施例で説明したルールを適用すると、図5の文書（a）、（b）から、図12の比較基準テーブルを参照することにより、図13の文書木（a）、（b）ができる。無視タグである＜章番号＞は、（ルール5）によって、ノードとして割り当てられていない。

【0022】手順204：文書木のノードを単位に差分抽出を行う。無視タグはノードとして存在しないため、比較されず、全体の差分抽出に影響を与えることはない。

手順205：一致しなかったノードのみ、今度は文字列単位で差分抽出を行う。手順204で比較しなかった無視タグおよびその中身もこの段階で比較を行う。

【0023】手順206：端末装置102に差分結果の表示を行う。図5の文書（a）との文書（b）の差分抽出を行った結果例を図14に示す。図14（b）は、図5（a）の構造化文書を基準として図5（b）の構造化文書との差分を取った場合の差分文字列の抽出結果であり、図14（a）は、図5（b）の構造化文書を基準として図5（a）の構造化文書との差分を取った場合の差

分文字列の抽出結果である。図5(a)の構造化文書を基準として図5(b)の構造化文書との差分を取り、図14(b)の差分文字列の抽出結果を得た場合について説明すると、手順204における文書木のノードを単位にした差分抽出では、図13(a)、(b)において、〈論文〉、〈／論文〉と、〈著者名〉平成太郎〈／著者名〉と、〈章〉構造化文書の差分抽出方式〈／章〉は一致と判断され、図14(b)では一致部分として表示されている。次に、手順205では、手順204で〈章〉構造化文書の差分抽出方式〈／章〉は一致と判断されているので、この一致部分に係る〈章番号〉、〈／章番号〉は一致と判断され、「第2章」は「第1章」とは一致しないので差分として抽出され、図14(b)のように表示される。また、手順204において、図13(b)の〈章〉構造化文書とは？〈／章〉は不一致と判断されるので、この〈章〉構造化文書とは？〈／章〉と、この不一致部分に係る〈章番号〉第1章〈／章番号〉は差分として抽出され、図14(b)のように表示される。

【0024】以上の手順により差分抽出を行うと、まず文書木のノード単位、すなわち構造単位で比較を行っているため、例えば、ノード1301と1302はこの時点で対応付けられる。よって、図6のような構造にまたがった対応付けは起こらないことが分かる。また、文書木のノード単位の比較では、無視タグの比較を行わないため、無視タグの中身の差異が全体の差分抽出に影響を与えないことが分かる。

【0025】(処理例3)実施例の具体的な処理例の3番目として、恒等タグおよび同等タグをもつ場合を図7の文書例で説明する。

手順201：文書編集プログラム104で、構造化文書の編集を行う。図7の(a)から図7の(b)を編集したとする。

【0026】手順202：比較対象であるSGML文書のDTDに対応した、比較基準テーブル107を読み込む。対応する比較基準テーブルが存在しない場合、テーブルの作成及び登録を行う。図7の例では、例えば、図15のような比較基準テーブルを作成する。すなわち、〈著者名〉を恒等タグとして定義する。この場合、タグ同士が一致しない限り、文字列同士を対応させない。また、〈項目〉と〈初項目〉を同等タグと定義する。この場合、〈項目〉と〈初項目〉は同じ論理構造とされる。

【0027】手順203：差分抽出プログラム106が呼び出されたら、比較基準テーブル107を参照しながら、SGML文書を構造化文書解析プログラム105によって解析し、文書木を作成する。実施例で説明したルールを適用すると、図7の文書(a)、(b)から、図15の比較基準テーブルを参照することにより、図16の文書木(a)、(b)ができる。図16中の1601、1602、1603は(ルール6)によって、同じ

タグ名に変換されている。

【0028】手順204：文書木のノードを単位に差分抽出を行う。同等タグは同じタグ名になっているため差分として抽出されない。

手順205：一致しなかったノードのみ、今度は文字単位で差分抽出を行う。

【0029】手順206：端末装置102に差分結果の表示を行う。図7の文書(a)と(b)の差分抽出を行った例を図17に示す。図17(b)は、図7(a)の構造化文書を基準として図7(b)の構造化文書との差分を取った場合の差分文字列の抽出結果であり、図17(a)は、図7(b)の構造化文書を基準として図7

(a)の構造化文書との差分を取った場合の差分文字列の抽出結果である。図7(a)の構造化文書を基準として図7(b)の構造化文書との差分を取り、図17

(b)の差分文字列の抽出結果を得た場合について説明すると、手順204における文書木のノードを単位にした差分抽出では、図16(a)、(b)において、〈論文〉、〈／論文〉と、〈著者名〉平成太郎〈／著者名〉と、〈項目〉構造化文書の差分抽出方式〈／項目〉は一致と判断され、図17(b)では一致部分として表示されている。次に、手順205では、手順204で〈項目〉構造化文書とは？〈／項目〉は不一致と判断されているので、この不一致部分について文字単位で差分抽出を行ない、〈項目〉構造化文書とは？〈／項目〉は差分として抽出され、図17(b)のように表示される。

【0030】以上の手順により差分抽出を行うと、タグ名が違っても、文書の論理構造が同じもの同士は、対応付けられることが分かる。

【0031】(処理例4)実施例の具体的な処理例の4番目として、比較禁止タグをもつ場合を図18の文書例で説明する。

手順201：文書編集プログラム104で、構造化文書の編集を行う。図18の(a)から図18の(b)を編集したとする。

【0032】手順202：比較対象であるSGML文書のDTDに対応した、比較基準テーブル107を読み込む。対応する比較基準テーブルが存在しない場合、テーブルの作成及び登録を行う。図18の例では、例えば、図19のような比較基準テーブルを作成する。すなわち、〈差出人〉と〈受取人〉とは比較禁止タグとする。この場合、〈差出人〉と〈受取人〉は中身を比較しない。

【0033】手順203：差分抽出プログラム106が呼び出されたら、比較基準テーブル107を参照しながら、SGML文書を構造化文書解析プログラム105によって解析し、文書木を作成する。実施例で説明したルールを適用すると、図18の文書(a)、(b)から、図19の比較基準テーブルを参照することにより、図20の文書木(a)、(b)ができる。

【0034】手順204：文書木のノードを単位に差分抽出を行う。〈差出人〉と〈受取人〉は比較するタグ同士が比較禁止タグなので、そのノード以下（子ノード）は比較しない。

手順205：一致しなかったノードのみ、今度は文字単位で差分抽出を行う。

【0035】手順206：端末装置102に差分結果の表示を行う。図18の文書(a)と(b)の差分抽出を行った例を図21に示す。図21(b)は、図18

(a)の構造化文書を基準として図18(b)の構造化文書との差分を取った場合の差分文字列の抽出結果であり、図21(a)は、図18(b)の構造化文書を基準として図18(a)の構造化文書との差分を取った場合の差分文字列の抽出結果である。図18(a)の構造化文書を基準として図18(b)の構造化文書との差分を取り、図21(b)の差分文字列の抽出結果を得た場合について説明すると、手順204における文書木のノードを単位にした差分抽出では、図18(a)、(b)において、〈メモ〉、〈／メモ〉は一致と判断され、〈受取人〉、〈／受取人〉とその中身である〈所属〉〇〇銀行〈／所属〉〈氏名〉平次太郎〈／氏名〉は、〈差出人〉と〈受取人〉とが比較禁止タグであるので差分とされ、〈本文〉こんにちは。お元気ですか？〈／本文〉は不一致と判断される。次に、手順205では、手順204で〈本文〉こんにちは。お元気ですか？〈／本文〉は不一致と判断されているので、この不一致部分について文字単位で差分抽出を行ない、「お元気ですか？」が差分として抽出される。この結果、図21(b)のように表示される。

【0036】以上の手順により差分抽出を行うと、中身を比較しないタグ同士を比較禁止タグとして登録しておけば、そのノード以下（子ノード）は比較されず、〈差出人〉と〈受取人〉の中身の所属や名前が対応付けられることがなく、編集者に対して、より適切な差分抽出結果を提示することが出来る。

【0037】

【発明の効果】構造化文書の論理構造に応じた比較基準を定義し、これを満たすよう差分を抽出することで、論理構造の意味に応じた、編集者の感覚に合う差分が抽出され、また、構造を表す文書木のノード単位で差分抽出を行い、一致しなかったノード同士を文字単位で差分を抽出することで、構造にまたがった差分も抽出されないため、編集者は論理構造にあった差分を把握することが出来、構造化文書の編集の効率が上がる。

【図面の簡単な説明】

【図1】本発明の実施例の構成を示す図である。

【図2】本発明の実施例の処理手順を示す図である。

【図3】構造化文書の第一の例を示す図である。

【図4】構造化文書の第一の例を従来の方式で差分抽出した結果例を示す図である。

【図5】構造化文書の第二の例を示す図である。

【図6】構造化文書の第二の例を従来の方式で差分抽出した結果例を示す図である。

【図7】構造化文書の第三の例を示す図である。

【図8】構造化文書の第三の例を従来の方式で差分抽出した結果例を示す図である。

【図9】構造化文書の第一の例に対する比較基準テーブルの例を示す図である。

【図10】構造化文書の第一の例から図9の比較基準テーブルに基づき作成した文書木を示す図である。

【図11】構造化文書の第一の例を図9の比較基準テーブルに基づき差分抽出した結果例を示す図である。

【図12】構造化文書の第二の例に対する比較基準テーブルの例を示す図である。

【図13】構造化文書の第二の例から図12の比較基準テーブルに基づき作成した文書木を示す図である。

【図14】構造化文書の第二の例を図12の比較基準テーブルに基づき差分抽出した結果例を示す図である。

【図15】構造化文書の第三の例に対する比較基準テーブルの例を示す図である。

【図16】構造化文書の第三の例から図15の比較基準テーブルに基づき作成した文書木を示す図である。

【図17】構造化文書の第三の例を図15の比較基準テーブルに基づき差分抽出した結果例を示す図である。

【図18】構造化文書の第四の例を示す図である。

【図19】構造化文書の第四の例に対する比較基準テーブルの例を示す図である。

【図20】構造化文書の第四の例から図19の比較基準テーブルに基づき作成した文書木を示す図である。

【図21】構造化文書の第四の例を図19の比較基準テーブルに基づき差分抽出した結果例を示す図である。

【符号の説明】

101 CPU

102 端末装置

103 記憶装置

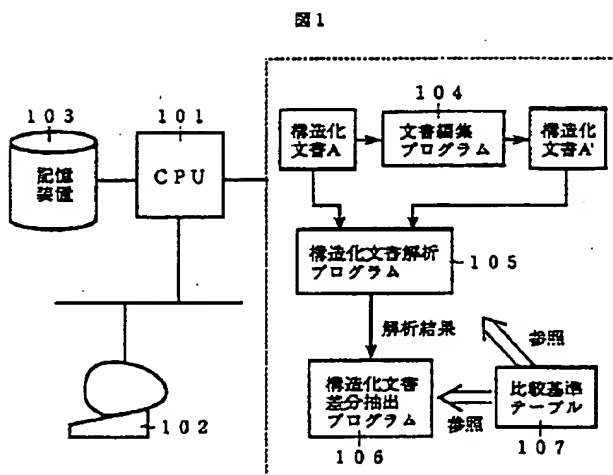
104 文書編集プログラム

105 構造化文書解析プログラム

106 構造化文書差分抽出プログラム

107 比較基準テーブル

【図1】



【図3】

図3

(a)

```

<メモ>
<氏名>平成太郎</氏名>
<本文>
  こんにちは。
</本文>
</メモ>
  
```

(a)

(b)

```

<メモ>
<発信日>平成6年11月20日</発信日>
<本文>
  こんにちは。お元気ですか?
</本文>
</メモ>
  
```

(b)

【図4】

図4

(a)

```

<メモ>
<氏名>平成太郎</氏名>
<本文>
  こんにちは。
</本文>
</メモ>
  
```

(a)

(b)

```

<メモ>
<発信日>平成6年11月20日</発信日>
<本文>
  こんにちは。お元気ですか?
</本文>
</メモ>
  
```

(b)

下線部：差分文字列

【図5】

図5

(a)

```

<論文>
<著者名>平成太郎</著者名>
<章>
  <章番号>第1章</章番号>
  構造化文書の差分抽出方式
</章>
</論文>
  
```

(a)

(b)

```

<論文>
<著者名>平成太郎</著者名>
<章>
  <章番号>第1章</章番号>
  構造化文書とは?
</章>
<章>
  <章番号>第2章</章番号>
  構造化文書の差分抽出方式
</章>
</論文>
  
```

(b)

【図9】

図9

比較基準テーブル		
項番	タグ	基準の種類
1	<氏名>	恒等タグ
2	<発信日>	恒等タグ

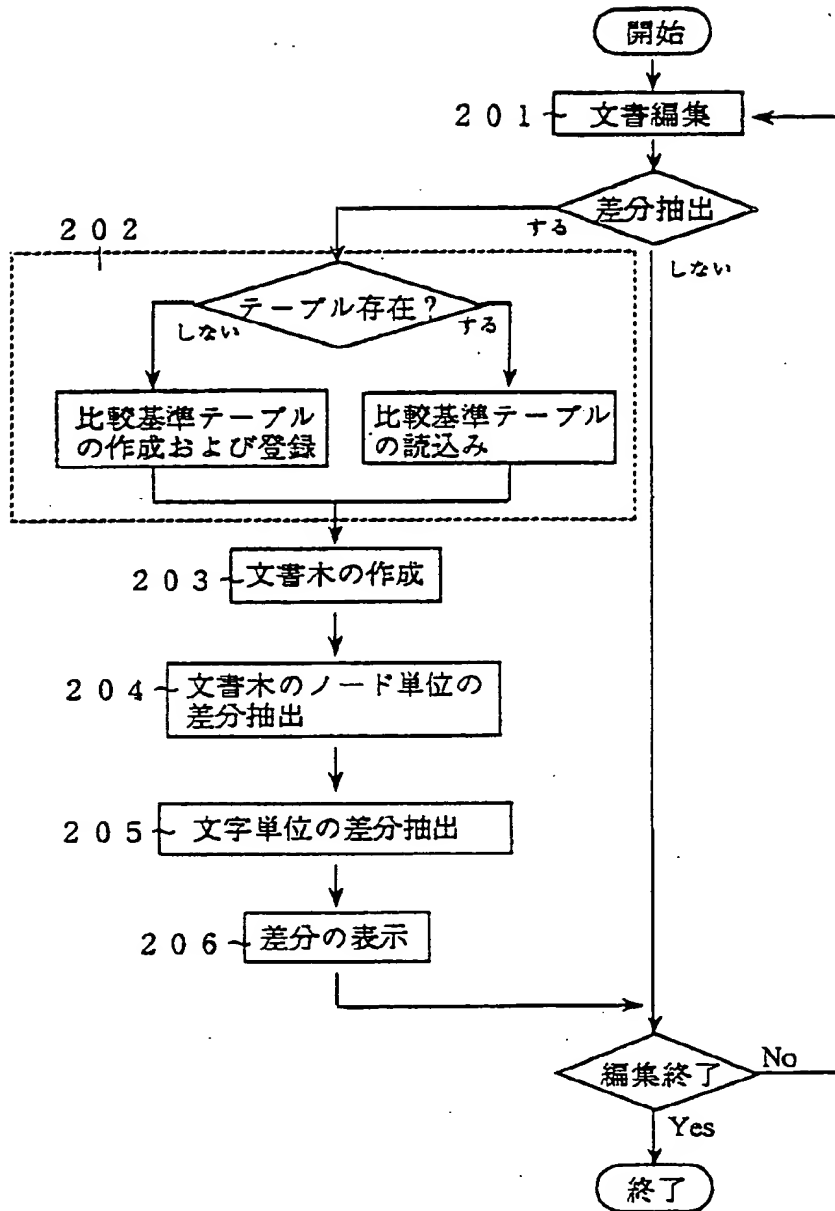
【図12】

図12

比較基準テーブル		
項番	タグ	基準の種類
1	<著者名>	恒等タグ
2	<章番号>	無視タグ

【図2】

図2



【図6】

図6

(a)

```

<論文>
<著者名>平成太郎</著者名>
<章>
<章番号>第1章</章番号>
構造化文書の差分抽出方式
</章>
</論文>
  
```

(a)

(b)

```

<論文>
<著者名>平成太郎</著者名>
<章>
<章番号>第1章</章番号>
構造化文書とは?
</章>
<章>
<章番号>第2章</章番号>
構造化文書の差分抽出方式
</章>
</論文>
  
```

(b)

下級部：差分文字列

【図7】

図7

(a)

```

<論文>
<著者名>平成太郎</著者名>
<初項目>
構造化文書の差分抽出方式
</初項目>
</論文>
  
```

(a)

(b)

```

<論文>
<著者名>平成太郎</著者名>
<初項目>
構造化文書とは?
</初項目>
<項目>
構造化文書の差分抽出方式
</項目>
</論文>
  
```

(b)

【図15】

図15

比較基準テーブル		
項番	タグ	基準の種類
1	<著者名>	恒等タグ
2	<項目> <初項目>	同等タグ

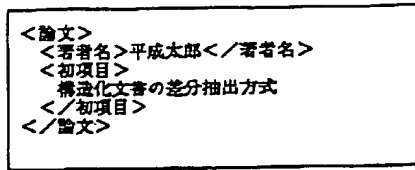
【図19】

図19

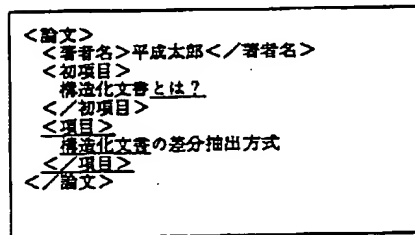
比較基準テーブル		
項番	タグ	基準の種類
1	<差出人> <受取人>	比較禁止タグ

【図8】

図8



(a)

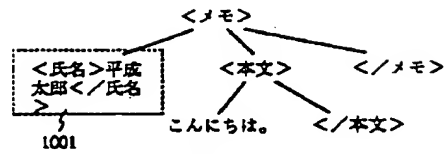


(b)

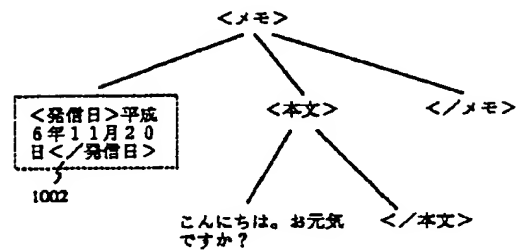
下線部：差分文字列

【図10】

図10



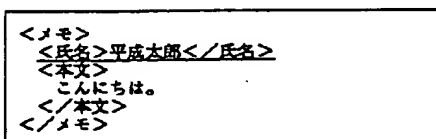
(a)



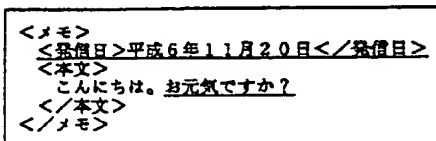
(b)

【図11】

図11



(a)

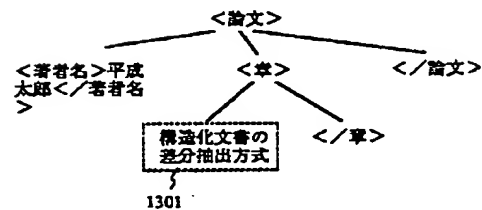


(b)

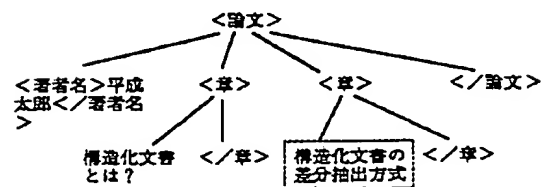
下線部：差分文字列

【図13】

図13



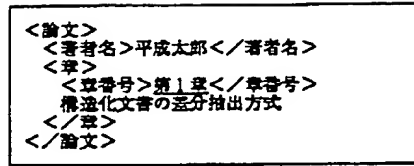
(a)



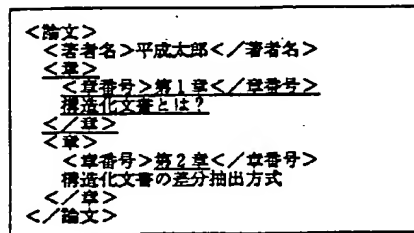
(b)

【図14】

図14



(a)

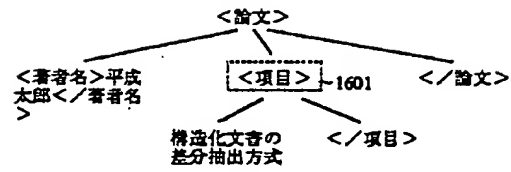


(b)

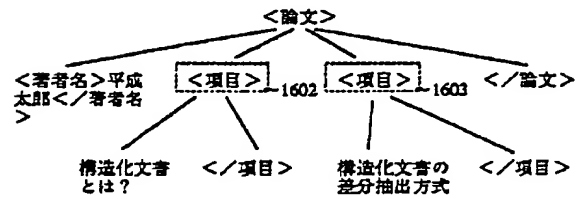
下線部：差分文字列

【図16】

図16



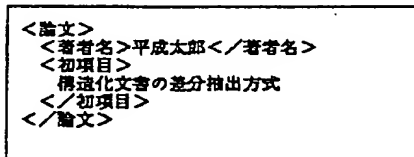
(a)



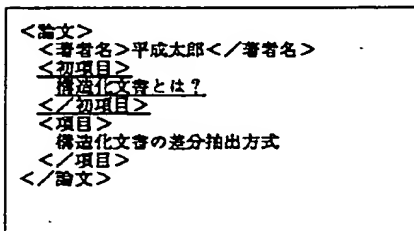
(b)

【図17】

図17



(a)

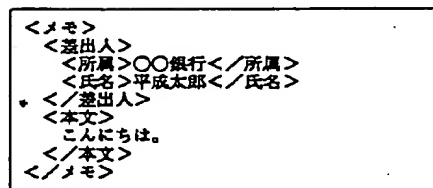


(b)

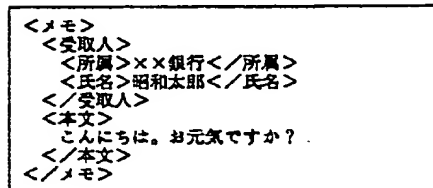
下線部：差分文字列

【図18】

図18

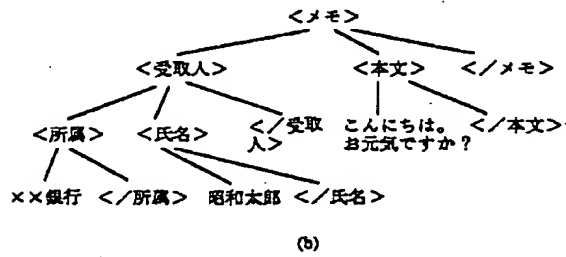
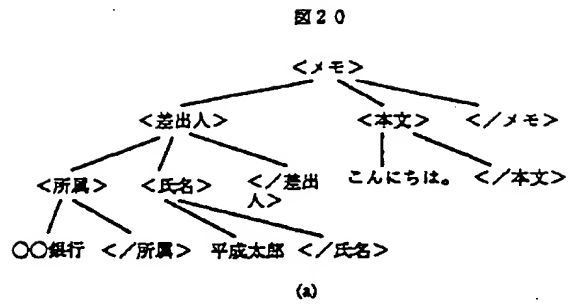


(a)



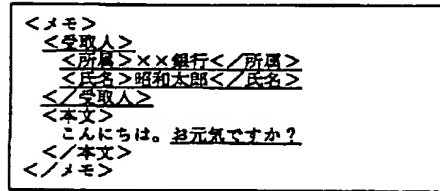
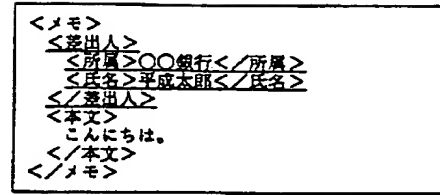
(b)

【図20】



【図21】

図21



下線部：差分文字列